图书情报工作
第65卷 第20期 2021年10月

HHa中心性算法:一种基于h指数和Ha指数的复杂网络节点排序算法*

■ 刘佳程[1,2] 马廷灿[1,2,3] 岳名亮[1,2,3]

[1] 中国科学院大学图书情报与档案管理系 北京 100190 [2] 中国科学院武汉文献情报中心 武汉 430071

[3] 科技大数据湖北省重点实验室 武汉 430071

摘 要:[目的/意义]针对复杂网络中的重要节点的识别,设计一种节点中心性算法,在传染病防控、舆情监控、产品营销、人才发现等方面发挥作用。[方法/过程]同时考虑节点的高影响力邻居的数量及其总体影响,提出HHa节点中心性算法,在真实网络和人工网络上,使用SIR传染病模型模拟信息传播过程,采用单调函数M和肯德尔相关系数作为评价指标验证HHa中心性算法的有效性、准确性以及稳定性。[结果/结论]实验表明,与7种经典的中心性算法相比,HHa中心性算法得出的排序结果M值为0.999等,排名第2;肯德尔系数为0.845等,高于其他算法0.15左右,排名第1且表现稳定。采用HHa中心性算法识别网络中的重要节点具备可行性。

关键词:复杂网络 节点中心性 节点影响力 h指数 HHa中心性算法

分类号:TP393.0 TP301.6

DOI:10.13266/j.issn.0252-3116.2021.20.010

1 引言

从自然世界到人类社会,复杂网络无处不在,各个系统中的个体和关系都可以抽象成具有节点和边的网络[1-3],如人类所处的社交网络、科研合作网络,宏观世界中的电力网络、交通网络等,微观世界中的蛋白质交互网络、基因网络、病毒传播网络等。网络科学的发展为人们认识客观世界增添了新的视角,帮助人们更好地理解关系的变化[3]、信息的传播[4-6]、传染的扩散[7-8]、疾病的治疗[9]等。

无论网络拥有何种结构和功能,关键节点都在信息传播中扮演着重要角色。因此,在复杂网络中识别有影响力的节点获得越来越多的关注,并且广泛应用于众多领域。比如,传染病将要爆发时,风险节点的精确查找直接影响到接种免疫政策的制定和后续防控的难易程度[10];推广新产品时,代言人物的正确选取和营销手段的精准策划能够快速创造商业价值[11];在谣言传播及舆情监控中,意见领袖的发言可以迅速遏制谣言并且引导积极的舆论走向[12];在科研合作网络中,及时发现和支持重要人才可以促进知识流动,增加学术交流[13]。

为此,本文将科学计量学中的h型指数运用到复杂网络分析中,并做出适当的改进,提出一种考虑节点高影响力邻居数量及总体质量的新型中心性算法,希望能做到计算复杂度和排序准确度的平衡,更加高效地发现复杂网络中高影响力的重要节点;然后在不同结构和功能的真实和人工网络上进行实验,验证该中心性算法的有效性、准确性以及稳定性。以期在传染病防控、舆情监控、产品营销、人才发现等方面发挥作用。

2 相关研究

识别网络中有影响力节点的中心性算法可以大致分为四类:第一类算法最简单直接,基于节点的邻居信息,如度中心性[14],节点i的度定义为与节点i直接相连的节点数目,即节点i的直接邻居数目。考察与节点直接相连的邻居数目,该中心性简单、直观,便于计算,认为节点的邻居数目越多,则节点越重要,但却忽

* 本文系2020年中国科学院文献情报能力建设专项"科技领域战略情报研究咨询体系建设"(项目编号:E0290001)研究成果之一。

作者简介:刘佳程(ORCID:0000-0001-5418-7307),硕士研究生;岳名亮(ORCID:0000-0002-1138-6661),副研究员,博士;马廷灿(ORCID:0000-0001-5985-384X),研究馆员,硕士,通讯作者,E-mail:matc@whlib.ac.cn。

收稿日期:2021-04-14 修回日期:2021-06-27 本文起止页码:92-100 本文责任编辑:杜杏叶

略了节点所处的环境信息，比如节点在网络的位置、邻居的质量。k-shell 中心性[15]根据节点在网络中的位置确定节点的重要程度，该中心性认为位于网络核心的节点即使其邻居数很少，重要性通常也很高，k-shell 中心性计算复杂度低，适用于大规模网络。但 k-shell 中心性得出的排序结果区分度不高，很多节点拥有同样的 k 核数，且 k-shell 在某些网络（如星型图、BA 人工网络）中不能发挥最优表现，且 k-shell 在逐层分解中考虑节点的邻居信息不够全面，仅考虑节点的剩余邻居数；第二类算法关注节点在网络中控制信息流的能力，即节点所处的路径，如在交通网、电力网中，重要节点往往扮演"桥"的角色，中介中心性[16]从网络中任一节点对的最短路径出发，通过一个节点的最短路径数越多，该节点的影响力越大。接近中心性[17]认为与网络中其他节点平均距离越小的节点越重要，即节点与网络中其他节点越接近，那么该节点对信息流动的影响也就越大，可以理解为接近中心性是借助信息在网络中的传播速度来界定节点的重要程度，但接近中心性只能用于连通网络。值得注意的是，中介中心性和接近中心性都是基于网络全局信息的中心性算法，它们要遍历整个网络才可以得到所有节点的影响力，时间复杂度较高；第三类算法是基于特征向量的中心性算法，特征向量中心性[18]方法是该类方法的典型代表，该方法判断节点的重要程度时，不仅考虑了节点的邻居数量，还考虑了邻居节点的重要性，但当网络中存在度较大节点时，特征向量中心性会将度大的节点列为重要节点，对度小的其他节点区分度不高；第四类算法是将科学计量学的相关指标运用到复杂网络中，用来衡量节点的影响力。比如 A. Korn 等参照科学计量学中的 h 指数，提出了 lobby-index（h 指数中心性）[19]，虽然计算简单，但排序结果的区分度并不高，即网络中的众多节点可能拥有相同的 h 指数。

本文提出的中心性算法属于第四类算法，因此接下来重点讨论以 h 指数为基础的重要节点挖掘方法。L. Y. Lü 等[20]揭示了度、h 指数和核数之间的内在关系，度、h 指数和核数，可以通过一个简单的算子 h 连接起来，而度、h 指数和核数是一连串作用的初态、中间态和稳态。Q. Liu 等[21]将节点的 h 指数和该节点所有邻居的 h 指数进行加和作为识别网络中重要节点的指标，但节点的邻居的度信息以及邻居的质量仍未完全考虑。A. Zareie 等[22]借鉴 h 指数的定义，设计了一个累计函数来计算节点的邻居度信息，并且设置参数将邻居的度信息进行加和以对节点影响力进行排序，该方法认为与高度节点相连的低度节点也可以作为高影响力节点的候选，但该指标涉及可调参数、计算复杂，而且准确率较经典算法提升不大。S. Zhao 等[23]和 L. Gao 等[24]将 h 指数应用于赋权网络，将边的权重考虑在内，定义了 h-degree 和 weighted h-index 用以衡量赋权网络中节点的重要性，但该类方法仍未克服 h 指数中心性原有的缺点。P. L. Yang 等[25]将最短距离和 h 指数相结合，求最短距离需要计算节点同网络中其他所有节点的距离，类似上文提到的接近中心性，需要遍历整个网络，计算复杂度高，并且只能求得高影响力的节点群，没有进一步对高影响力节点的影响力大小进行区分。Y. X. Li 等[26]基于 h 指数和节点的 r 阶和 n 阶邻居的 h 指数提出了衡量网络中一组节点影响力的中心性（h-index group centrality），但 r、n 以及阈值条件需要根据不同结构的网络进行调整，对网络的适用程度不高，且计算量较大。卢鹏丽等[27]提出了将节点的度和 h 指数结合的节点中心性度量方法，但该种度量方法使用时需要对涉及的参数进行调试，并且未考虑节点邻居的质量信息。A. Abbasi 等在 2013 年提出了节点的 al-index 指数，并将其定义为对该节点 h 指数有贡献的邻居节点的平均度[28]，单纯使用该种中心性分辨重要节点的准确性不高，且并未将其与 h 指数等联合用于节点中心性排序。

通过上述分析，可以看出，现有各种节点中心性算法虽各有优势，但仍有可改进之处：①目前的中心性算法有着"简单却不够准确，或者准确但太复杂"的问题，如何在简单和准确中把握平衡应是新的节点中心性算法考虑的关键点，要确保在计算复杂度低的情况下考虑尽可能多的节点信息；②网络结构和功能复杂多样，节点中心性算法是否在多种网络保持有效性和准确性，应在大量真实网络和人工网络上进行实验，确保节点中心性算法的普适性。

针对上述问题，本文将科学计量学中的 h 型指数迁移运用到复杂网络分析中，进行节点中心性排序时，在考虑节点的高影响力邻居的数量的同时，也考虑高影响力邻居的总体影响，提出了一种新的中心性算法。为了评价该中心性算法的有效性、准确性以及稳定性，应用 SIR 传染病模型模拟信息传播过程，将肯德尔系数作为算法准确性评价指标，实验结果表明，本文提出的中心性算法能够产生区分度很高的排序结果，并且与其他 7 种经典中心性算法相比，该中心性算法得到的排序结果更加准确，针对不同结构和功能的网络，该中心性算法具有更好的稳定性。本文提出的中心性算



法可进一步用于真实场景，有望在优秀人才选拔、传染病爆发防控、舆情信息监控、产品营销推广等方面开发实际产品。

3　HHa 中心性及其计算方法

如图 1 所示，用网络 $G(V,E)$ 表示一个有 V 个顶点，E 条边的网络。节点 v 的度定义为与节点 v 直接相连的邻居数量，用 $d(v)$ 表示节点 v 的度，$N(v)$ 表示 v 的邻居，$I(v)$ 为 v 的邻居中度大于或等于 v 的 h 指数的节点。h 指数在网络中的定义：节点 v 的 h 指数是 h，就说明这个节点有 h 个邻居，它们的度都不小于 h，而且这个节点的其他所有邻居的度都不大于 h，用 $h(v)$ 表示。本文定义节点 v 的 Ha 指数：在节点 v 的邻居中，将度大于或等于 $h(v)$ 的节点的度进行加和，所求得的数便是节点 v 的 Ha 指数，用 $Ha(v)$ 表示，$Ha(v)$ 的计算方式如公式(1)所示：

$$Ha(v) = \sum_{w \in I(v)} d(w) \qquad \text{公式(1)}$$

图 1　示例网络

易得节点 v_1 的 h 指数为 3，按照公式(1)计算节点 v_1 的 Ha 指数为 13，将 h 指数和 Ha 指数一同使用，考虑节点的高影响力邻居的数量和总体质量，定义求节点 HHa 中心性的算法：利用 h 指数作为第一关键字对网络中各节点进行降序排列，再将各节点的 Ha 指数作为第二关键字进行降序排列。可以看出，节点的 h 指数和 Ha 指数越高，则该节点的影响力越大。HHa 中心性无须遍历整个网络就可获取网络中节点的有关信息，算法时间复杂度低，且计算简单。例如，在图 1 中，用 $HHa(v)$ 表示节点 v 的 HHa 指数，节点 v_1、v_2、v_3 的 HHa 指数如公式(2) – 公式(4)所示：

$$HHa(v_1) = (h(v_1),\ Ha(v_1)) = (3, 3+4+3+3) = (3,13) \qquad \text{公式(2)}$$

$$HHa(v_2) = (h(v_2),\ Ha(v_2)) = (3,\ 4+3+4) = (3,11) \qquad \text{公式(3)}$$

$$HHa(v_3) = (h(v_3),\ Ha(v_3)) = (2,\ 6+4) = (2,10) \qquad \text{公式(4)}$$

因此，$HHa(v_1) > HHa(v_2) > HHa(v_3)$，即节点 v_1、v_2、v_3 的影响力和重要程度依次递减。

4　HHa 中心性算法的有效性和准确性检验

4.1　数据集

为了检验 HHA 中心性算法的表现，本文将使用不同规模的 8 个真实网络，包括：Karate[29]、Dolphins[30]、Jazz[31]、Usair97[32]、Email[33]、Netscience[34]、Yeast[35]、Power[36]。其中 Karate、Dolphins、Yeast、Power 是无向无权网络，Usair97、Jazz、Netscience 是无向加权网络，Email 是有向无权网络。相关网络的节点数、边数、平均度、传播阈值如表 1 所示：

表 1　实验数据集基本信息

网络	节点数	边数	平均度	传播阈值
Karate	34	78	4.588	0.129
Dolphins	62	159	5.129	0.147
Jazz	198	2 742	27.697	0.026
USAir97	332	2 126	12.807	0.023
Email	1 133	5 451	9.622	0.054
Netscience	1 589	2 742	3.451	0.144
Yeast	2 361	7 182	6.084	0.060
Power	4 941	6 594	2.669	0.258

4.2　评估方法

本文使用单调函数 M[37] 来量化不同中心性算法给出的排序结果的单调性，即该中心性算法是否能对各个节点的影响能力进行差异化识别，单调函数 M 的定义如公式(5)所示：

$$M(R) = \left(1 - \frac{\sum_{r \in R} n_r(n_r - 1)}{n(n-1)}\right)^2 \qquad \text{公式(5)}$$

其中，n 是排序表 R 的元素个数，n_r 是排序表 R 上具有相同排名的元素个数。单调函数 M 量化了排序表中具有相同名次元素的比例。如果排序表 R 是完全单调的，则单调性 $M(R)$ 为 1；如果排序表中所有节点具有相同的排名，则单调性 $M(R)$ 为 0。

目前，易感 – 感染 – 恢复(SIR)传播模型[38]被广泛用于解释和模拟流行病或信息的传播过程[15,21 – 22,24 – 27,37,41]。因此，在本文中，为了评估排序方法的性能，采用 SIR 传播模型来检验排序节点的真实影响力。在 SIR 模型中，假设要获得节点 v 的真实影

响力，在初始阶段将节点 v 设置为感染状态，其他节点全部设置为易感状态。在之后的每个步骤中，每个状态为感染的节点尝试以感染概率 β 感染其状态为易感的邻居，并且在每一步中每个被感染的节点以 γ 的概率转变为恢复状态。重复此过程，直到网络中没有状态为感染的节点为止。将传播过程结束时恢复状态的节点数作为估算最初感染节点 v 的影响的指标。节点 v 的真实影响力定义为在足够大的模拟传播次数后恢复节点数的平均值。在本文中，将模拟传播次数设置为 100，即 100 次 SIR 传播后的恢复节点数的平均值作为节点的真实影响力。为了量化各类中心性算法的准确性，本文采用 Kendall 相关系数[39] 作为排名相关系数。假设有两个排序结果 $X=(x_1, x_2, \cdots, x_n)$ 和 $Y=(y_1, y_2, \cdots, y_n)$，则两个排序结果对应元素形成的元素对集合为 $U=\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，对于该集合中的任意两个元素对 (x_i, y_i) 与 (x_j, y_j)，如果 $x_i > x_j$ 且 $y_i > y_j$ 或者 $x_i < x_j$ 且 $y_i < y_j$ 时，那么这两个元素是一致的；如果 $x_i > x_j$ 且 $y_i < y_j$ 或者 $x_i < x_j$ 且 $y_i > y_j$ 时，那么这两个元素是不一致的；如果 $x_i = x_j$ 或者 $y_i = y_j$ 时，这两个元素既不是一致的也不是不一致的。Kendall 相关系数计算方法如下：$T(X, Y) = (n_c - n_d)/0.5n(n-1)$。其中，$n_c$ 和 n_d 分别是一致元素对和不一致元素对的数量，n 为排序表 X 或 Y 包含的排序结果数量。

在本文中，基准排序结果为通过模拟 SIR 传播得到的网络中节点的真实影响力排序，另一个排序结果是通过节点中心性算法得到的网络中节点中心性排序结果。Kendall 系数的取值范围为[－1，1]，系数越接近于 1，两个排序结果越接近，说明该中心性算法得到的排序结果与模拟 SIR 传播得到的节点真实影响力排序结果越接近，准确性越高。

在实验过程中用 DC 表示度中心性、BC 表示中介中心性、CC 表示接近中心性、KS 表示 k-shell 中心性、EV 表示特征向量中心性、H 表示 h 指数中心性、Ha 表示 Ha 指数中心性、HHa 表示 HHa 中心性。

4.3 单调性检验

针对 8 个真实网络数据集，使用各类中心性算法得到相应的节点影响力排序结果，并使用单调函数 M 计算各个排序结果的单调性，各中心性算法排序结果的单调函数 M 值如表 2 所示。从表 2 可以看出，由于 HHa 中心性是基于节点的 h 指数和 Ha 指数得到的，所以其单调 M 值较 h 指数中心性的单调 M 值有很大提升，并且 HHa 中心性的单调 M 值与其他 7 种中心性单调 M 值相比，处于上游水平，仅次于特征向量中心性或中介中心性，排名第 2 或第 3。说明本文提出的 HHa 中心性可以较好地识别和区分网络中具有不同影响力的节点，各节点能够被赋予不同的指标值，本文算法具有竞争优势。

表 2　不同中心性算法在 8 个真实网络数据集上的单调函数 M 值

网络	$M(DC)$	$M(BC)$	$M(CC)$	$M(EV)$	$M(KS)$	$M(H)$	$M(Ha)$	$M(HHa)$
Karate	0.708	0.775	0.899	0.958	0.496	0.577	0.930	0.940
Dolphins	0.831	0.962	0.974	0.998	0.377	0.684	0.957	0.974
Jazz	0.966	0.989	0.988	0.999	0.794	0.938	0.998	0.999
Usair97	0.859	0.697	0.989	0.995	0.811	0.836	0.992	0.994
Email	0.887	0.940	0.999	1.000	0.809	0.858	0.994	0.998
Netscience	0.738	0.090	0.909	0.917	0.700	0.707	0.897	0.908
Yeast	0.734	0.701	0.996	0.997	0.674	0.698	0.984	0.990
Power	0.593	0.832	1.000	1.000	0.246	0.393	0.881	0.924

4.4 准确性检验

接下来，通过将不同中心性算法得到的排序结果与从 SIR 模拟传播获得的节点影响力排序表进行比较，计算两者之间的 Kendall 相关系数，研究不同中心性算法的准确性，Kendall 系数越大，表明相应的中心性算法得到的节点影响力排序结果准确性越高。在 SIR 模拟中，首先将感染概率 β 设置在网络的传播阈值[40] $\beta_{th} = <k>/<k^2>$ 附近，其中 $<k>$ 为网络的平均度，如果感染率 β 远小于传播阈值，则传播过程无法正常进行；如果感染率 β 远远大于传播阈值，则传播过程会在网络中迅速激烈地展开，并导致所有节点都被感染，最后都变为恢复状态。因此，将感染率 β 设置在传播阈值附近是合理和可行的。其次，设置节点从感染状态到恢复状态的恢复概率为 $\gamma = 1/<k>$[41]。与传统的 SIR 传播模型相比，本文感染率和恢复率的个性化设置更能体现网络的实际情况，能够更有效地评价网络中节点的真实影响力。各中心性算法得出的排序结果与 SIR 传播模型排序结果的 Kendall 系数如表 3

所示。从表 3 可以看出,在 8 个真实网络中,HHa 中心性的 Kendall 相关系数均大于 h 指数中心性的 Kendall 相关系数;在 Karate、Dolphins、Usair97、Email 和 Yeast 网络中,HHa 中心性的 Kendall 相关系数较其他 7 种中心性算法处于首位,准确性最高;在 Jazz 网络中,仅次于度中心性或特征向量中心性,识别效果处于第二位;在 Power 网络中,仅次于特征向量中心性和 Ha 指数中心性,识别效果位于第三位;在 Netscience 网络中,识别效果位列第四。

表 3 不同节点中心性算法在 8 个真实网络数据集上与节点真实影响力的肯德尔相关系数

网络	β_{th}	β	$T(DC)$	$T(BC)$	$T(CC)$	$T(EV)$	$T(KS)$	$T(H)$	$T(Ha)$	$T(HHa)$
Karate	0.129	0.130	0.640	0.569	0.611	0.647	0.569	0.579	0.572	**0.697**
Dolphins	0.147	0.150	0.689	0.523	0.571	0.575	0.538	0.664	0.703	**0.711**
Jazz	0.026	0.026	**0.595**	0.354	0.507	0.543	0.542	0.586	0.576	0.587
Usair97	0.023	0.023	0.760	0.521	0.727	0.801	0.773	0.776	0.797	**0.834**
Email	0.054	0.055	0.824	0.667	0.706	0.702	0.796	0.819	0.782	**0.829**
Netscience	0.144	0.145	0.644	0.188	**0.863**	0.849	0.614	0.624	0.802	0.713
Yeast	0.059	0.060	0.730	0.576	0.754	0.712	0.744	0.750	0.790	**0.845**
Power	0.258	0.260	0.389	0.299	0.392	0.595	0.315	0.345	0.598	0.542

为了测试感染概率 β 的影响,在 SIR 传播模型中设置 β 的值为 0.01 - 0.20。如图 2 所示,在 Dolphins、Usair97 和 Yeast 网中,随着感染概率 β 的增大,HHa 的肯德尔相关系数 T 值都显著高于其他 7 种中心性,在 Email 网中 HHa 中心性的肯德尔 T 值在感染概率 β 较小时表现较好,能够明显看出识别准确性处于首位,但随着感染率 β 增大,HHa 中心性的优势不再明显,但仍处于前列。可以推测,网络的结构特性会影响信息在整个网络上的传播过程。因此,本文接下来使用人工网络研究网络结构所带来的影响。

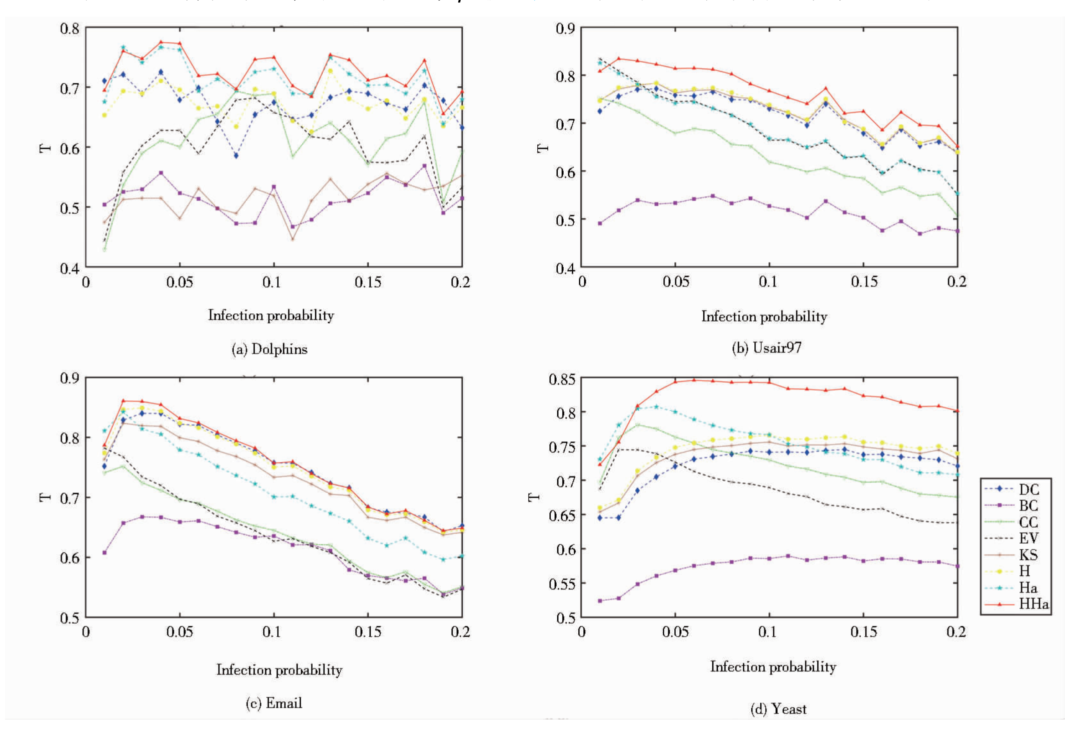


图 2 真实网络中不同感染率下的节点影响力与各节点中心性算法的肯德尔相关系数

本文采用 LFR 人工网络[42]作为网络实验模型。在 LFR 网络中,有众多可以表征网络结构特征的参数,通过设置不同的参数取值,可以构建不同结构的网络,以此来看网络结构的变化是否会对 HHa 的适用性产生影响。设置初始默认参数为:节点数量为 1 000,平均度为 5,最大度为 50,网络中节点的度符合参数为 $\gamma=2$ 的指数分布,社团的混合参数 $\mu=0.2$。通过变化节点数量、平均度、度分布指数、以及社团混合参数来评估这些参数变化所带来的影响。

首先,设置网络节点数量分别为 $N=500$,1 000,

1 500,如图 3 所示,当感染率较小时,接近中心性、特征向量中心性和 Ha 指数中心性是衡量节点重要程度的有效方法,但当感染率增大时,接近中心性、特征向量中心性以及 Ha 指数中心性的表现迅速下降,HHa 中心性较其他中心性有显著优势。

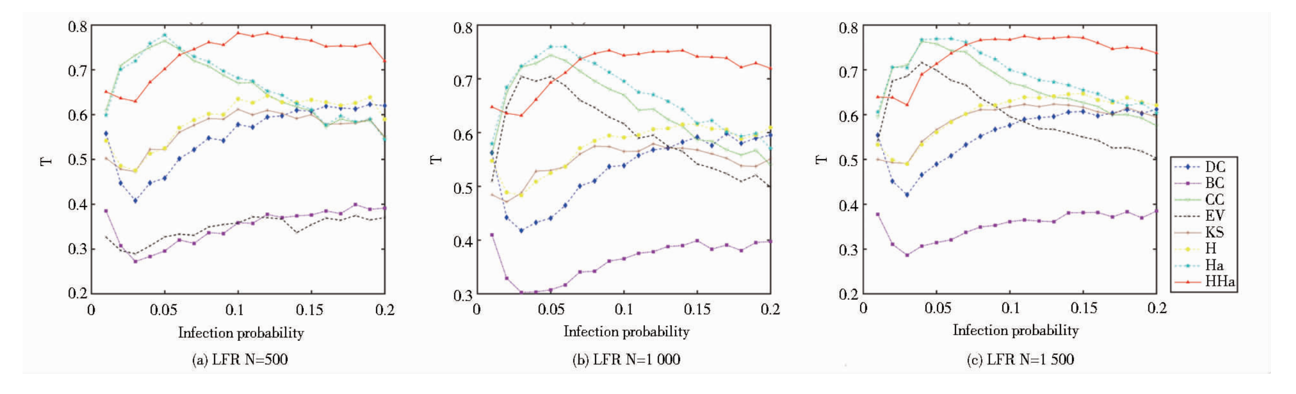


图 3　LFR 网络(改变网络中的节点数量)中各节点中心性算法与真实影响力的肯德尔系数变化

其次,设置网络平均度分别为 $<k> = 5,10,15$,如图 4 所示,值得注意的是,随着网络平均度的增大,HHa 中心性逐渐接近度中心性,甚至当平均度 $<k> = 15$ 时,8 种中心性表现趋于重叠,并且随着感染率的增大,8 种中心性算法的准确性都明显下降。这可能是由于当网络平均度较大和感染概率较大时,选取网络的任一节点作为初始感染态节点都能使得信息迅速且广泛地传播,节点的重要程度在这种情况下被削弱。但 HHa 中心性算法仍能在感染率较小时获得最优表现。

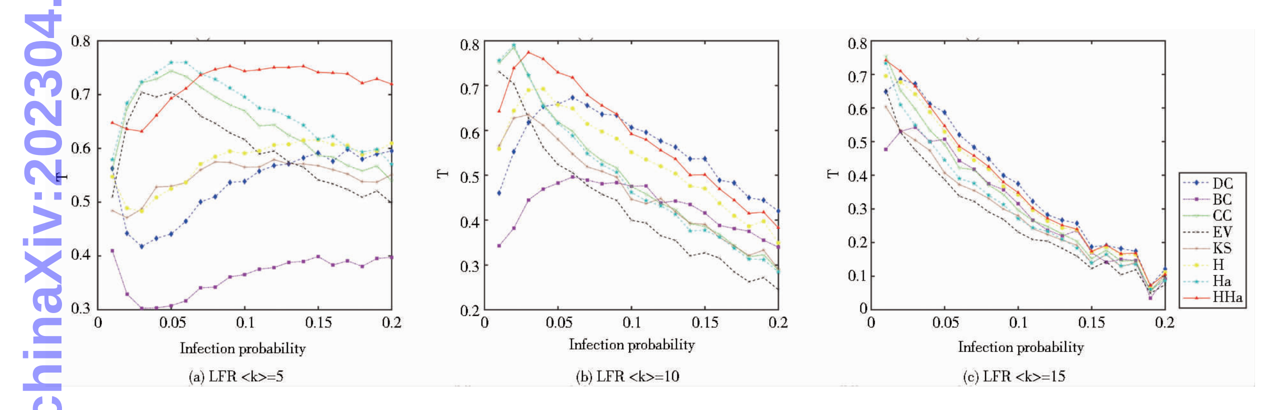


图 4　LFR 网络(改变网络的平均度)中各节点中心性算法与真实影响力的肯德尔系数变化

接下来,设置节点的度分布指数 $t = 2, 2.5, 3$,如图 5 所示,可以看出,无论度分布指数如何取值,HHa 中心性算法表现都比较平稳,虽然在感染率较小时只能处于中上游水平,但随着感染率的增加,HHa 中心性算法的表现总是会超过其他中心性算法,居于首位。

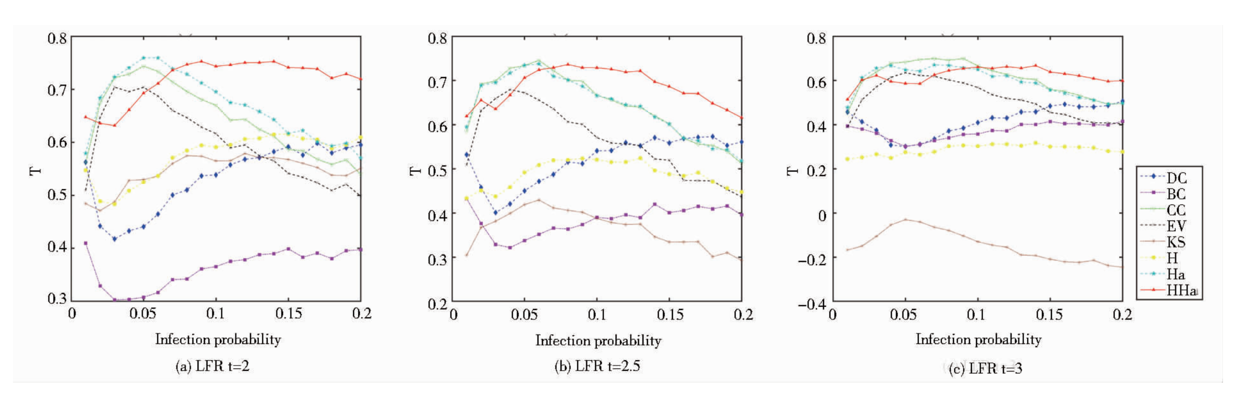


图 5　LFR 网络(改变网络的度分布指数)中各节点中心性算法与真实影响力的肯德尔系数变化

最后,设置社团混合参数 $\mu=0.2,0.5,0.8$,如图6所示。社团结构是网络中非常重要的一个特性,当社团混合参数较低时,LFR benchmark 会创建具有明显社团结构的随机网络。实验结果表明,无论社团结构清晰与否,HHa 中心性算法总能显著地识别出网络中具有影响力的节点,并且表现远远高于其他7种中心性算法。

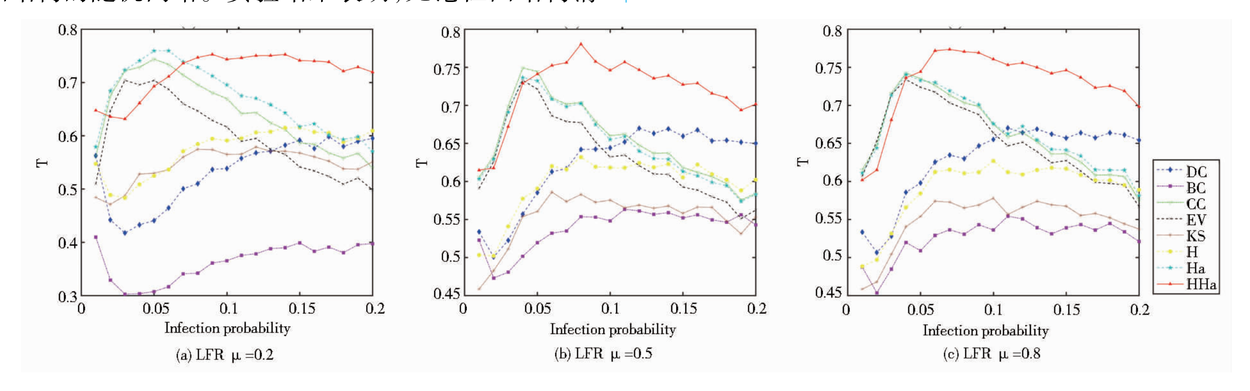


图6 LFR 网络(改变网络的社团混合参数)中各节点中心性算法与真实影响力的肯德尔系数变化

总的来说,在基于网络规模节点数或其他参数变化下的 LFR 网络上,虽然网络的拓扑性质发生了变化,但 HHa 中心性算法总能更准确地识别出网络中的重要节点。当感染率比较小时,由于传播难以扩散,传播范围只限于局部邻居节点,这时拥有邻居越多的节点通常能影响更多的节点,因此感染率较低时,度中心性算法表现较好。而当感染率较大时,邻居之外的节点被感染的概率会增大,度中心性等算法的不足逐渐显现,HHa 中心性算法的优势变得更加明显。相比其他中心性算法,HHa 中心性算法的稳定性更高,即对参数变化的敏感性更低,在拥有不同网络结构特征的 LFR 网络中,验证了 HHa 中心性算法的有效性和准确性。

5 总结与讨论

本文提出了一种有效的排序方法——HHa 中心性算法,即使用 h 指数和 Ha 指数来衡量复杂网络中节点的影响力。目前现有的算法存在"简单而不准确,准确但太复杂"的缺点,本文提出的 HHa 中心性算法将科学计量学里的指标迁移运用到复杂网络中,在进行节点中心性排序时,同时考虑了节点的高影响力邻居的数量和高影响力邻居的总体影响,较好地把握了算法的简单性和准确性之间的平衡。实验结果证明,HHa 中心性算法是一种可以评估网络信息传播能力、识别网络中的重要节点的简单且有力的方法。首先,本文通过单调函数 M 来衡量各中心性能否区分节点的重要性大小,在8个真实网络上进行实验,实验结果表明 HHa 中心性较 h 指数中心性和 Ha 指数中心性能产生更多的单调排名,和其他7种指标相比处于中上游地位。此外,为了度量 HHa 中心性得到的节点影响力排序的准确性,将 SIR 传染病模型的传播结果作为节点的真实影响力,使用肯德尔相关系数来衡量各中心性算法得到的排序结果和 SIR 传染病模型得到的排序结果的一致性,与其他中心性算法相比,HHa 中心性算法准确性最高。基于 LFR benchmark,在具有不同拓扑性质的人工网络上进行实验,HHa 中心性算法的准确性和稳定性明显优于其他中心性方法算法。

本文的实验网络集包含有向\无向、加权\无权网络,HHa 中心性在这些网络上都取得了很好的效果,说明 HHa 中心性应用范围广泛,但由于 HHa 中心性是基于 h 指数中心性和 Ha 指数中心性得到的,而 h 指数中心性和 Ha 中心性不考虑网络中边的方向及权重,所以 HHa 中心性也未将边的方向及权重考虑在内,今后可以基于这两点对 HHa 中心性进行改进,让其考虑的网络信息更多,在确保应用普适性的同时,提高其应用针对性和选择性。值得注意的是,节点的重要性是在网络的结构和功能中体现出来的,因此评价中心性算法的有效性和准确性必须要结合网络的具体结构和功能进行,虽然本文在 LFR 网络上对提出的 HHa 中心性算法的适用性和一般性进行了较完整的测试,但本文选取的真实网络数据集仍不够多,仍未涵盖所有的功能目标。

此外,SIR 模型只是对现实世界的高度模拟和抽象的结果,虽然具有一定的指导意义,但由于真实网络中的个体特性不一、不确定性较大,得出的结果可能和真实情况有出入。

最后,本文提出的 HHa 中心性算法在充分进行小规模的真实实验后,可扩大应用范围,在实际场景中进

行使用,辅助人工高效率地进行重要节点的发现。比如发现著名学者和领军人才、获取领域重要论文、查找微博高影响力用户、识别关键致病基因、提取文章关键词、控制传染病流行、搜寻罪犯及恐怖分子、挖掘意见领袖等。

参考文献:

[1] BARABÀSI A L. Network science[M]. Cambridge:Cambridge University Press, 2016.
[2] STROGATZ S H. Exploring complex networks[J]. Nature, 2001, 410(6825):268 - 276.
[3] NEWMAN M. The structure and function of complex networks[J]. Siam review, 2003,45(2):167 - 256.
[4] BORGE-HOLTHOEFER J, MORENO Y. Absence of influential spreaders in rumor dynamics[J]. Physical review e, 2012, 85(2):026116.
[5] HOSNI A, LI K, AHMAD S. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors[J]. Information sciences, 2020, 512:1458 - 1480.
[6] AHMED W, VIDAL-ALABALL J, DOWNING J, et al. COVID-19 and the 5G conspiracy theory: social network analysis of twitter data[J]. Journal of medical intertnet research, 2020, 22(5):e19458.
[7] COLIZZA V, BARRAT A, BARTHÉLEMY M, et al. The modeling of global epidemics: stochastic dynamics and predictability[J]. Bulletin of mathematical biology,2006, 68(8):1893 - 1921.
[8] WANG W, TANG M, EUGENE S H,et al. Unification of theoretical approaches for epidemic spreading on complex networks[J]. Reports on progress in physics. physical society (Great Britain), 2017,80(3):036603.
[9] LIU X R, HONG Z Y, LIU J, et al. Computational methods for identifying the critical nodes in biological networks[J] Briefings in bioinformatics, 2020,21(2):486 - 497.
[10] 刘振杰,赵姝,陈洁,等. 一种新的基于节点重要性的免疫策略研究[J]. 南京大学学报(自然科学),2017,53(2):350 - 356.
[11] LÜ L Y, ZHANG Y C, YEUNG C H,et al. Leaders in social networks, the delicious case[J]. Plos one, 2011, 6(6):e21202.
[12] 陈芬, 付希, 何源,等. 融合社会网络分析与影响力扩散模型的微博意见领袖发现研究[J]. 数据分析与知识发现, 2018, 2(12): 60 - 67.
[13] YOU X M, MA Y H, LIU Z Y, et al. Representation method of cooperative social network features based on node2vec model[J]. Computer communications,2021,173:21 - 26.
[14] FREEMAN L C. Centrality in social networks: conceptual clarification[J]. Social network, 1979, 1(3):215 - 239.
[15] KITSAK M, GALLOS L K, HAVLIN S, et al. Identification of influential spreaders in complex networks[J]. Nature physics, 2010,6(11):888 - 893.
[16] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1):35 - 41.
[17] SABIDUSSI G. The centrality index of a graph[J]. Psychometrika, 1966, 31(4):581 - 603.
[18] BONACICH P. Factoring and weighting approaches to status scores and clique identification[J]. Journal of mathematical sociology, 1972, 2(1):113 - 120.
[19] KORN A, SCHUBERT A, TELCS A. Lobby index in networks[J]. Physica a: statistical mechanics and its applications, 2009, 388(11):2221 - 2226.
[20] LÜ L Y, ZHOU T, ZHANG Q M, et al. The h-index of a network node and its relation to degree and coreness[J]. Nature communications, 2016, 7:10168.
[21] LIU Q, ZHU Y X, JIA Y,et al. Leveraging local h-index to identify and rank influential spreaders in networks[J]. Physica a: statistical mechanics and its applications, 2018,512: 379 - 391.
[22] ZAREIE A,SHEIKHAHMADI A. Ehc: extended h-index centrality measure for identification of users' spreading influence in complex networks[J]. Physica a: statistical mechanics and its applications, 2019, 514:141 - 155.
[23] ZHAO S, ROUSSEAU R, YE F Y. H-degree as a basic measure in weighted networks[J]. Journal of informetrics, 2011, 5(4):668 - 677.
[24] GAO L, YU S,LI M,et al. Weighted h-index for identifying influential spreaders[J]. Symmetry. 2019, 11(10):1263.
[25] YANG P L, LIU X, XU G Q. An extended clustering method using h-index and minimum distance for searching multiple key spreaders[J]. International journal of modern physics c, 2019, 30(7): 1940008.
[26] LI Y X, SHENG Y Q, YE X Z. Group centrality algorithms based on the h-index for identifying influential nodes in large-scale networks[J]. International journal of innovative computing, information and control,2020,16(4):1183 - 1201.
[27] 卢鹏丽,于洲. 基于度与 H 指数扩展的复杂网络节点排序方法[J]. 兰州理工大学学报,2020,46(5):100 - 106.
[28] ABBASI A. H-type hybrid centrality measures for weighted networks J]. Scientometrics,2013,96(2):633 - 640.
[29] ZACHARY W W. An information flow model for conflict and fission in small groups[J]. Journal of anthropological research, 1977, 33(4): 452 - 473.
[30] LUSSEAU D, SCHNEIDER K, BOISSEAU O J, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations[J]. Behavioral ecology & sociobiology, 2003, 54(4):396 - 405.
[31] GLEISER P M, DANON L. Community structure in jazz[J]. Advances in complex systems, 2003, 6(4):565 - 573.
[32] BATAGELJ V,MRVAR A. Pajek datasets[DB/OL]. [2021 - 04 - 07]. http://vlado. fmf. uni-lj. si/pub/networks/data.
[33] GUIMERÀ R, DANON L, DIAZ-GUILERA A, et al. Self-similar community structure in a network of human interactions[J]. Physical review e, 2003, 68(6):065103.

[34] NEWMAN M. Finding community structure in networks using the eigenvectors of matrices[J]. Physical review e, 2006, 74(3): 036104.
[35] BU D, ZHAO Y, CAI L, et al. Topological structure analysis of the protein-protein interaction network in budding yeast[J]. Nucleic acids research, 2003(9):2443-2450.
[36] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks. [J]. Nature, 1998, 393(6684):440-442.
[37] BAE J, KIM S, et al. Identifying and ranking influential spreaders in complex networks by neighborhood coreness[J]. Physica a: statistical mechanics and its applications, 2014, 395(4):549-559.
[38] KERMACK W O, MCKENDRICK A G. A contribution to the mathematical theory of epidemics[J]. Proceedings of the royal society a mathematical physical & engineering sciences, 1927, 115(772): 700-721.
[39] KENDALL M. A new measure of rank correlation[J]. Biometrika, 1938, 30(1/2):81-93.
[40] CASTELLANO C, PASTOR-SATORRAS R. Thresholds for epidemic spreading in networks[J]. Physical review letters, 2010, 105(21):218701.
[41] 邹青，张莹莹，陈一帆，等. 社交网络中一种快速精确的节点影响力排序算法[J]. 计算机工程与科学，2014，36(12):2346-2354.
[42] LANCICHINETTI A, FORTUNATO S, RADICCHI F. Benchmark graphs for testing community detection algorithms[J]. Physical review e, 2008, 78(4):046110.

作者贡献说明：

刘佳程：实验设计、数据处理与论文起草；
马廷灿：实验设计与论文修订；
岳名亮：论文修订。

HHa Centrality Algorithm: A Node Centrality Algorithm Based on the H-Index and Ha-Index

Liu Jiacheng[1,2] Ma Tingcan[1,2,3] Yue Mingliang[1,2,3]
[1] Department of Library, Information and Archives Management, University of Chinese Academy of Sciences, Beijing 100190
[2] Wuhan Library of Chinese Academy of Sciences, Wuhan 430071
[3] Hubei Key Laboratory of Big Data in Science and Technology, Wuhan 430071

Abstract: [**Purpose/significance**] For the identification of important nodes in complex networks, the paper designs a node centrality algorithm, which plays an important role in infectious disease prevention and control, public opinion monitoring, product marketing, talent discovery and so on. [**Method/process**] This paper proposed a new node centrality algorithm, the HHa node centrality algorithm, taking both the number of the node's high influence neighbors and their total influence into consideration. On the real network and artificial network, the Susceptible-Infected-Recovered (SIR) model was used to simulate the information dissemination process, and the monotonic function M and Kendall correlation coefficient were used as evaluation indicators to verify the effectiveness, accuracy and stability of the HHa centrality algorithm. [**Result/conclusion**] The experimental results show that, compared with the 7 classic centrality algorithms, the HHa centrality algorithm ranks 2nd with a monotonic result of 0.999, and the Kendall coefficient is 0.845, which is higher than other algorithms' accuracy about 0.15, ranking 1st and performing robustly. It is feasible to use HHa centrality algorithm to identify important nodes in the network.

Keywords: complex networks node centrality node influence h-index HHa centrality algorithm